JOURNAL OF GEOPHYSICAL RESEARCH, VOL. xx, NO. xx, PAGES 1?? , FEBRUARY 2002

Kalman Filter Chemical Data Assimilation: A Case Study in January 1992

D. J. Lary
Data Assimilation Office, NASA Goddard Space Flight Centre, Greenbelt, MD

B. Khattatov
NCAR, Boulder, CO

H. Mussa
Department of Chemistry, University of Cambridge, England

Abstract. This paper describes a Kalman filter chemical data assimilation system and its use for analysing a vertical atmospheric profile during January 1992. The vertical profile was at an equivalent PV latitude (ϕ_e) of 55°S and consisted of 21 potential temperature (θ) levels spaced equally in $\log(\theta)$ between 400 K and 2000 K. This equivalent latitude was chosen as it was well observed during January 1992 by instruments on board the Upper Atmosphere Research Satellite (UARS).

1. Introduction

Measurements of atmospheric constituents made over the last decade or more have cost many millions of dollars/pounds to make. The intelligent use of this data on a wide variety of species is a non-trivial task as the observations are not co-located in time or space. Satellites make measurements of atmospheric constituents by a range of methods, and at a range of times and locations. The measurements are not made on a regular spatial grid or at the same times of day. Since the analysis of satellite measurement is so complex, the measurements have not been used to their full potential.

In comparison to the analysis of meteorological variables, chemical trace species has received little attention. Current methods tend to be either simple comparisons of observations with a model (which are not necessarily constrained to be directly comparable) and/or treat species independently, ignoring the complex balances which exist between species. Moreover, the large diurnal variations in the concentrations of many species are either accounted for in very simple ways, or avoided by analysing concentrations at fixed local time. This is a great shame as the shape of a species diurnal cycle, and the relative partitioning between species, contains a lot of valuable information that is completely wasted if we do not use a technique such that can exploit this information. Naturally such information can only be exploited if it includes a theoretical understanding of the chemical system. Data assimilation is a valuable assistant in making better use of observations of atmospheric chemistry. This paper describes a Kalman filter for chemical data assimilation with observation quality control and analyses skill assessment cast in flow-tracking coordinates.

Paper number 2002JAxx.
0148-0227/02/2002JAxx89.00

2. Flow-Tracking Coordinate System

We want to look at the detailed interactions between chemical species and exploit the propagation of information between chemical species by using a Kalman filter which calculates the time evolution of the full co-variance matrix. This is expensive and so as a first step we will take a lagrangian approach. To give us global analyses we then use a two-dimensional array of independent time evolving chemical box models (described in section 3). This two-dimensional array is arranged in an equivalent-PV latitude theta flow tracking co-ordinate system [*Lary et al.*, 1995a]. This approximation is certainly valid for our analysis interval of one day, and often for up to ten days. It is a way of largely separating the effects of chemistry and dynamics. Because a major component of the variability of trace gases is due to the atmospheric motions we use a co-ordinate system to perform our data assimilation that 'moves' with the large scale flow pattern.

In addition, the Kalman filter chemical data assimilation is computationally expensive, one diurnal cycle for one vertical profile taking 35 minutes of computer time on a 1.7 GHz intel pentium IV computer (this includes the first guess, assimilation, and analyses run). So it is useful to have a global 2D assimilation by using an equivalent PV latitude (ϕ_e), potential temperature (θ) co-ordinate system. Our grid has 21 potential temperature levels spaced equally in $\log(\theta)$ between 400 K and 2000 K, and 32 equivalent PV latitudes spaced evenly between -90° and 90°. Here we consider in detail just one of these 32 profiles, the one at 55°S as it was well observed during January 1992 by instruments aboard the Upper Atmosphere Research Satellite (UARS).

3. Chemical Scheme

We use the extensively validated AutoChem model described [*Fisher and Lary*, 1995; *Lary et al.*, 1995b; *Lary*, 1996]. The model is explicit and uses an adaptive-timestep,

error monitoring time integration scheme for stiff systems of equations [*Stoer and Bulirsch*, 1980; *Press et al.*, 1992]. AutoChem was the first model to ever have the facility to perform 4D variational data assimilation (4D-VAR) [*Fisher and Lary*, 1995] and now also includes a Kalman filter [*Khattatov et al.*, 1999]. AutoChem uses kinetic data largely based on [*DeMore et al.*, 1997] and [*Atkinson et al.*, 1997].

Our usual chemical system contains a total of 60 species. 55 species are time integrated, namely: $O(^1D)$, $O(^3P)$, O_3, N, NO, NO_2, NO_3, N_2O_5, HONO, HNO_3, HO_2NO_2, CN, NCO, HCN, Cl, Cl_2, ClO, ClOO, OClO, Cl_2O_2, $ClNO_2$, ClONO, $ClONO_2$, HCl, HOCl, CH_3OCl, Br, Br_2, BrO, $BrONO_2$, BrONO, HBr, HOBr, BrCl, H_2, H, OH, HO_2, H_2O_2, CH_3, CH_3O, CH_3O_2, CH_3OH, CH_3OOH, CH_3ONO_2, $CH_3O_2NO_2$, HCO, HCHO, CH_4, CH_3Br, CF_2Cl_2, CO, N_2O, CO_2, H_2O. The remaining 5 species are not integrated and not in photochemical equilibrium, namely: O_2, N_2, $HCl_{(S)}$, $H_2O_{(S)}$, $HNO_{3(S)}$. The model contains a total of 420 reactions, 278 bimolecular reactions, 32 trimolecular reactions, 60 photolysis reactions, 4 cosmic ray processes, 46 heterogeneous reactions.

3.1. Radiative Transfer Calculations

A key part of the chemical model is the calculation of photolysis rates. In this study photolysis rates are calculated using full spherical geometry and multiple scattering [*Lary and Pyle*, 1991a, b; *Meier et al.*, 1982; *Nicolet et al.*, 1982] with a treatment of spherical geometry [*Anderson*, 1983]. The photolysis rate used for each time step is obtained by ten point Gaussian-Legendre integration [*Press et al.*, 1992]. These calculations are updated on every assimilation iteration to ensure that the improved ozone profile at a given equivalent latitude is used to calculate the photolysis rates.

4. Quality Control

Observation quality control is a central part of chemical data assimilation. Our system transforms the observations into a flow tracking coordinate system. We then use many observations to produce a single pseudo observation profile. We then deal with a complete pseudo observation profile at a time.

No observation is used unless the ratio of the observational error, σ, to the observed concentration, χ, which we will call the quality ratio, Q_r, does not exceed a certain specified threshold (1).

$$Q_r = \frac{\sigma}{\chi} \tag{1}$$

The value of this threshold is specified for each observed species separately based on the characteristics of the instrument involved. This criteria has proved to be important in removing rouge observations and is recommended to others engaged in chemical data assimilation. In addition there is option that an observation is only used if with its associated uncertainty it overlaps the current analysis concentration and its associated uncertainty (this criteria is not always applied as it can lead to an incestuous relationship between the observations chosen and the analyses). We usually have two iterations of the Kalman filter, the first using all observations that passed the quality ratio test, and the second where the analyses state is also used to perform quality control.

4.1. Generation of the Pseudo Observations

Generating a pseudo observation is necessary for the assimilation as both the observation and analyses need to be dealing with the same location. In addition, it allows an improved signal to noise as many observations are used in forming just one psuedo observation.

We have chosen to deal with a profile of pseudo observations at a time as the two quality control criteria mentioned above can lead to gaps in our vertical profile. These are easy to fill in by eye, but for an algorithm to deal with the data voids we need to consider an entire profile at a time in our flow tracking co-ordinate space. In a full 3D assimilation the 3D co-variance matrix would be performing this task. However, in this study we are using a full Kalman filter with a detailed chemistry. To make this computationally achievable we use multiple 0D box models which are stacked into a series of profiles. Which, as mentioned earlier, then gives us a 2D global assimilation with 21 potential temperature levels spaced equally in $\log(\theta)$ between 400 K and 2000 K, and 32 equivalent PV latitudes spaced evenly between -90° and 90°.

The key point about our generation of pseudo observations is that we deal with the ratio of the observed concentration, χ_O, to the analysis concentration, χ_A, which we will call the observed concentration ratio, R, not with the observed concentrations directly (2).

$$R = \frac{\chi_O}{\chi_A} \tag{2}$$

Where the analysis concentration is interpolated to the location of each observation in turn. We then look at the many observation points that fall between the bottom of a given grid box and the top of the current grid box as a distribution of observed concentration ratios. We then take the observed concentration ratio for that grid box as the median observed concentration ratio of this distribution of observed concentration ratios. The median is used as it is not affected by any large outliers in the distribution of observed concentration ratios. If there are more than a threshold number of observations, N, (usually 10) then we just take the median of the N most accurate observations. The pseudo observation for a grid box, χ_{pseudo}, is then simply the product of the median observed concentration ratio, R_{median}, and the current analysis concentration, χ_A (3).

$$\chi_{pseudo} = R_{median}\chi_A \tag{3}$$

If we have any gaps in the vertical pseudo observation profile we simply linearly interpolate the median observed concentration ratio from the available points above and below the gap. Since this ratio is generally quite close to one the interpolation is rather good, and better than performing an interpolation in concentration units. The concentration can change by more than an order of magnitude over the profile and contain strong gradients. In contrast R is generally close to one and does not contain strong gradients.

4.2. Observation Uncertainties

The uncertainty of the pseudo observation has two components. First, the observational uncertainty which is taken to be the median observed concentration uncertainty of the distribution of observed concentration uncertainties. Second, the representativeness which is taken to be the average

deviation of the distribution of observed concentrations for the grid box.

$$\sigma_{rep} = ADev(\chi_1 \ldots \chi_N) = \frac{1}{N}\sum_{j=1}^{N} |\chi_j - \bar{\chi}| \quad (4)$$

The average deviation, or mean absolute deviation, is a robust estimator of the width of the distribution [*Press et al.*, 1992].

It has been found desirable to include a moving average smoothing which involves the current grid box, and the boxes above and below.

The usefulness of generating pseudo observations in this way is particularly noticeable for those observational datasets that have gaps and those that are rather noisy.

4.3. Checking the Errors

It often hard to know if the observation and apriori errors have been correctly specified. In this particular study the best characterised error is the representativeness error. The model error growth is taken to be 5% per time step. The observation errors are taken directly from the values specified with the observed concentrations by the retrieval teams. However, these observational uncertainties could be in error as was found by *Menard et al.* [2000]; *Menard and Chang* [2000]. A useful check is

$$< (O-F)^2 > \approx \sigma_o^2 + \sigma_f^2 \quad (5)$$

5. Kalman Filter

The chemical Kalman filter [*Khattatov et al.*, 1999] allows one to optimally combine model simulations and measurements taking into account their respective uncertainties. Consider a model of a physical system represented by operator (generally nonlinear) $\mathcal{M}$, and let vector $\mathbf{x}$ with dimension N_x be a set of input parameters for the model. These input parameters are used to predict the state of the system, vector $\mathbf{y}$ with dimension N_y:

$$\mathbf{y} = \mathcal{M}(\mathbf{x}) \quad (6)$$

Assume that vector $\mathbf{x}$ represents the state of a time-dependent numerical photochemical model, i.e., concentrations of modeled species at model grid points in the atmosphere. In the case of a box model that includes N species, the dimension of vector $\mathbf{x}$ would be N. We will now limit the discussion to the case when $\mathcal{M}$ is used to predict the state of the system at some future time from past state estimates. Formally, in this case

$$\mathbf{x} = \mathbf{x_t} \quad , \quad \mathbf{y} = \mathbf{x_{t+\Delta t}} \quad (7)$$

$$\text{and } \mathbf{x_{t+\Delta t}} = \mathcal{M}(t, \mathbf{x_t}) \quad (8)$$

Let vector $\mathbf{y_o}$ contain observations of the state. Usually, the dimension of $\mathbf{y_o}$ is less than N_y, the dimension of the model space, since not all model species are usually observed. The connection between $\mathbf{y_o}$ and $\mathbf{y}$ can be established through the so-called observational operator $\mathcal{H}$:

$$\mathbf{y_o} = \mathcal{H}(\mathbf{x}) \quad (9)$$

Combining the above two equations, we get

$$\mathbf{y_o} = \mathcal{H}(\mathcal{M}(\mathbf{x})) \quad (10)$$

We now assume that that the probability density functions associated with $\mathbf{x}$ and $\mathbf{y}$ can be satisfactory approximated by Gaussian functions:

$$PDF(\mathbf{y}) \sim \exp\left(-\frac{(\mathbf{x}-\mathbf{x_t})^T \mathcal{C}^{-1} (\mathbf{x}-\mathbf{x_t})}{2}\right) \quad (11)$$

where $\mathbf{x_t}$ is the true value of $\mathbf{x}$ and $\mathcal{C}$ is the corresponding error covariance matrix. Its diagonal elements are the uncertainties (standard deviations) of $\mathbf{x}$, and the off-diagonal elements represent correlation between uncertainties of different elements of vector $\mathbf{x}$. The covariance matrix $\mathcal{C}$ is defined as

$$\mathcal{C} = \langle (\mathbf{x}-\mathbf{x_t})(\mathbf{x}-\mathbf{x_t})^T \rangle \quad (12)$$

where angle brackets represent averaging over all available realizations of x.

For most practical applications we need to introduce the linear approximation. In the linear approximation we assume that for small perturbations of the parameter vector $\Delta\mathbf{x}$ the following is approximately true:

$$\mathcal{M}(\mathbf{x}+\Delta\mathbf{x}) = \mathcal{M}(\mathbf{x}) + \mathcal{L}\Delta\mathbf{x} \quad (13)$$

Formally, $\mathcal{L}$ is a derivative of $\mathcal{M}$ with respect to x:

$$\mathcal{L} = \frac{d\mathcal{M}}{d\mathbf{x}} \quad (14)$$

For small variations of $\mathbf{x}$ one can show that the evolution of error covariance matrix $\mathcal{C}_t$ is given by:

$$\mathcal{C}_{t+\Delta t} = \mathcal{L}\mathcal{C}_t\mathcal{L}^T + \mathcal{Q} \quad (15)$$

Matrix $\mathcal{Q}$ is the error covariance matrix introduced to take into account uncertainties of the model calculations. The Kalman filter equations are

$$\mathbf{x_{t+\Delta t}} = \mathcal{M}(t, \mathbf{x_t})$$
$$\mathcal{C}_{t+\Delta t} = \mathcal{L}\mathcal{C}_t\mathcal{L}^T + \mathcal{Q}$$

$$\hat{\mathbf{x}}_\mathbf{t} = \mathbf{x_t} + \mathcal{C}_t\mathcal{H}^T(\mathcal{H}\mathcal{C}_t\mathcal{H}^T + \mathcal{O})^{-1}(\mathbf{y_o} - \mathcal{H}\mathbf{x_t}) \quad (16)$$

$$\hat{\mathcal{C}}_t = \mathcal{C}_t + \mathcal{C}_t\mathcal{H}^T(\mathcal{H}\mathcal{C}_t\mathcal{H}^T + \mathcal{O})^{-1}\mathcal{H}\mathcal{C}_t \quad (17)$$

At the end of each analysis period the model value ($\mathbf{x_t}$) and the corresponding observation ($\mathbf{y_o}$) are 'mixed' (see (16)) with weights inversely proportional to their respective errors to produce the analysis, $\hat{\mathbf{x}}_\mathbf{t}$. Then the model is integrated forward in time starting from the obtained analysis. Once an observation has been incorporated in the model, the analysis error covariance should be updated to reflect this (see (17)). In the absence of observations, the model state is updated using (8), while evolution of the error covariance is obtained from the linearized model equations as in (15).

If no observations are available, then

$$\hat{\mathbf{x}}_\mathbf{t} = \mathbf{x_t} \quad (18)$$

$$\hat{\mathcal{C}}_t = \mathcal{C}_t \quad (19)$$

6. Skill Scores

Once the data assimilation analyses has been performed we need to quantify how good the analyses is. This is done by generating a wide range of statistics. These statistics compare the observations used in making the analyses with the analyses itself. These statistics are presented in a web site automatically created by our software.

The diagnostics/statistics are as follows:

1. **Observation Increment** The difference between the first guess and the observations, also known as observed-minus-background differences or the innovation vector [*Daly*, 1991]. This is probably the best measure of forecast skill.

2. **Analysis Increment** The difference between the first guess and the final analyses, also known as analysis-minus-background differences or the correction vector [*Daly*, 1991]. This is a good measure of model bias.

3. **Cost** (jargon for accumulated difference between analyses and observations), both globally and for each single location in the analyses.

4. **Scatter plots** of observations against analyses, these can visually highlight any biases present.

5. **Normal Probability Plots of (Observation - Analysis) Values** are useful graphs for assessing whether data comes from a normal distribution.

6. **Quantile-Quantile Plots of observations against analyses**. A quantile-quantile plot is useful for determining whether two samples come from the same distribution (whether normally distributed or not). The quantile-quantile plot has three graphical elements. The pluses are the quantiles of each sample. The number of pluses is the number of data values in the smaller sample. The solid line joins the 25^{th} and 75^{th} percentiles of the samples. The dashed line extends the solid line to the extent of the sample. Figure 2 shows quantile-quantile plots for the analyses and observations from the assimilation presented here.

7. **Mean Error** (ME), **Bias, or Analysis-Observations (A-O)**, both globally and for each single location in the analyses.

$$ME = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k) \tag{20}$$

where o_k denotes the kth observation (or psuedo observation) and y_k the corresponding value from the analyses. This is a useful measure of the bias between the observations and analyses. Figure 1 shows examples of bias vertical profiles for the January 1992 test case considered here. Typically the bias is an order of magnitude less than concentrations and within the analyses error.

8. **Global Histograms of (Observation - Analysis) Values** The difference between the first guess and the final analyses.

9. **Mean Absolute Error** (MAE), both globally and for each single location in the analyses.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - o_k| \tag{21}$$

10. **Mean Square Error** (MSE), both globally and for each single location in the analyses.

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{22}$$

11. **Root Mean Square Error** (RMSE), both globally and for each single location in the analyses.

$$RMSE = \sqrt{MSE} \tag{23}$$

Figure 1 shows that, as would be expected, in many cases the bias is anticorrelated with the analysis increment. In other words, the assimilation process is trying to correct the bias that exists between the observations and model. This is just why the bias and analysis increments are most useful statistics in accessing the quality of both the model and observations.

6.1. A Cautionary Note

Data assimilation can easily cause a serious violation of conservation of mass if total mass, or total reactive nitrogen (NO_y= N + NO + NO_2+ $2N_2O_5$+ HONO + HNO_3+ HO_2NO_2+ CH_3ONO_2 + $CH_3O_2NO_2$ + $ClNO_2$ + ClONO + $ClONO_2$), chlorine (ClO_y= Cl + $2Cl_2$+ ClO + $2Cl_2O_2$+ $ClNO_2$ + ClONO + $ClONO_2$+ HCl + HOCl + CH_3OCl + BrCl), bromine (BrO_y= Br + $2Br_2$+ BrO + BrONO + $BrONO_2$+ HBr + HOBr + BrCl), or hydrogen (H_y= $2H_2$+ $2H_2O$+ $4CH_4$) are not included as control variables. To overcome this at the start of each timestep we note the NO_y, ClO_y, BrO_y and H_y, perform the assimilation, and then renormalise the NO_y, ClO_y, BrO_y and H_y. If this is not done totally unrealistic analyses can easily result.

An example of mass conservation affecting the analyses can be seen in Figure 1 where there is a noticeable bias in the H_2O analyses. This is because the total H_y is known quite accurately, and consequently the available observations of H_2O and CH_4 can not simultaneously be correct. This difference between the observations and analyses is highlighted in the H_2O quantile-quantile plot shown in Figure 2, and in the vertical profiles of bias, O-F, and A-F shown in Figure 1. Therefore the analyses have conserved mass by slightly reducing the levels of H_2O and CH_4, as can be seen in Figures 1 and 2. The adjustment in CH_4 is less than that in H_2O as CH_4 has the lower observation uncertainty.

A similar situation is found in the partitioning of reactive nitrogen. If we examine the vertical profiles of NO, NO_2, N_2O_5, HNO_3 and $ClONO_2$ in Figure 1 we see that between 10 and 30 mb there is a bias in all these species. For NO_2, N_2O_5, HNO_3 and $ClONO_2$ the analyses values are all less than the observations. This is because the total NO_y in this region is accurately known and the sum of the observed NO_2, N_2O_5, HNO_3 and $ClONO_2$ observations would considerably exceed the known NO_y. Consequently, the analyses has slightly reduced the concentrations of NO_2, N_2O_5, HNO_3 and $ClONO_2$ to ensure NO_y conservation. In the case of $ClONO_2$ it is very likely that there is an observational bias as $ClONO_2$ is also a significant component of the ClO_y family for which we also have HCl observations from HALOE, and between 10 and 30 mb there is no significant bias in HCl. Therefore the assimilated $ClONO_2$ concentrations, which have a bias relative to the observed $ClONO_2$ concentrations, are consistent with the observed and analysed HCl concentrations.

7. A Case Study

Let us now consider a case study from January 1992 where the Kalman filter chemical data assimilation system

described above was used to analyse a vertical atmospheric profile. The vertical profile was at an equivalent PV latitude (ϕ_e) of 55°S and consisted of 21 potential temperature (θ) levels spaced equally in $\log(\theta)$ between 400 K and 2000 K. This equivalent latitude was chosen as it was well observed during January 1992 by the the Halogen Occultation Experiment (HALOE), the Microwave Limb Sounder (MLS), and the Cryogenic Limb Array Etalon Sounder (CLAES) aboard the Upper Atmosphere Research Satellite (UARS). In addition, considering just one vertical profile allows a detailed examination of the diurnal cycle in species such as NO and NO_2.

Figure 1 shows vertical profiles of the chemical analyses produced by data assimilation for O_3, N_2O, NO, NO_2, N_2O_5, HNO_3, $ClONO_2$, HCl, H_2O and CH_4 overlaid with their pseudo observations. Shown on the same horizontal scale are vertical profiles of the representativeness error, the observation error and the analyses error. In a separate panel there are vertical profiles of the bias between the analyses and the pseudo observations, the analysis increment, (O-F), and (A-F). These quantities are all defined in Section 6 above.

Several points are noteworthy. As would be expected, the analyses error is normally always less than the combination of the observation and representativeness error (Figure 1). The only exception to this is for some parts of the NO and NO_2 vertical profiles. In each case it occurs above 35 km where there is an inconsistency between the observed NO, NO_2, and O_3 and the theoretical knowledge encapsulated in the model. In this region the photochemical theory of NO_x is well known and such an inconsistency did not occur when ATMOS data was used. It is therefore very likely that there is a bias in either, or both, the CLAES NO_2 observations or the HALOE NO observations. Due to the close chemical coupling between O_3 and NO_x this has also led to the largest ozone bias occurring above 40 km. The assimilation has combined all the information available and highlighted the inconsistency by the larger (O-F) and (A-F) values and by increasing the analyses uncertainty.

The biases in NO_2, N_2O_5, HNO_3 and $ClONO_2$ between 10 and 30 mb to ensure NO_y conservation have already been considered in Section 6.1 above.

Figure 2 shows quantile-quantile plots of observations against analyses. A quantile-quantile plot is useful for determining whether two samples come from the same distribution (whether normally distributed or not). The quantile-quantile plot has three graphical elements. The pluses are the quantiles of each sample. The number of pluses is the number of data values in the smaller sample. The solid line joins the 25^{th} and 75^{th} percentiles of the samples. The dashed line extends the solid line to the extent of the sample.

The quantile-quantile plots for O_3, NO, N_2O, HCl and CH_4 all show a good straight line relationship. This means that the shape of the observation and analyses probability distribution functions (PDFs) are the same to a very good approximation. The quantile-quantile plots for $ClONO_2$ shows a disagreement in the dotted line region, i.e. on the wings of the plot beyond the 75^{th} percentile, and the plots for NO_2, HNO_3 and H_2O show some minor discrepancies in the solid line region, these relate to the conservation issues mentioned above. The quantile-quantile plots for N_2O_5 shows the biggest discrepancy, the cause of this can be seen in Figure 1 where there is a large bias between the analyses and observations, again related to the conservation issues mentioned above.

The left hand column of plots in Figures 3 and 4 shows one diurnal cycle of the chemical analyses produced by data assimilation for a vertical profile at an equivalent PV latitude (ϕ_e) of 55°S consisting of 21 potential temperature (θ) levels spaced equally in $\log(\theta)$ between 400 K and 2000 K overlaid with the raw observations. The right hand column shows the corresponding analyses uncertainty overlaid with the observational uncertainty. As would be expected, the analyses uncertainty is less than the observational uncertainty as information propagates between variables and also comes from our apriori and theoretical description of the system.

It is noteworthy to see how the time variation in the analyses uncertainty is very different from species to species. Some species have very little change in their uncertainty, whereas species such as NO and NO_2 have a strong diurnal cycle in their uncertainty. Yet other species without a significant diurnal cycle, such as HCl, are affected by using observations of NO and NO_2. This shows the propagation of information between species within data assimilation and can be seen clearly in Figures 3 and 4. For example, the uncertainty of NO and HCl (right hand column of figures) are both affected by the observations of NO_2.

8. Summary

This paper gives a detailed description of a Kalman filter chemical data assimilation system, and an example of its use from January 1992. The system is designed to aid in the analysis and quality control of atmospheric observations made by remote sensing and in-situ instruments. Quality control has been found to be an essential part of the assimilation.

The assimilation has performed well and highlighted likely inconsistencies (biases) in the NO, NO_2, N_2O_5, HNO_3 and $ClONO_2$ observations between 10 and 30 mb, in O_3 and NO_x above 40 km, and in H_2O and CH_4 throughout much of the stratosphere. Such inconsistencies were not encountered when using high quality ATMOS data and thus show the value of chemical data assimilation as part of the validation of remotely sensed chemical data. We hope to use this system in the validation of ENVISAT data.

Acknowledgments. It is a pleasure to acknowledge: The Royal Society for a Royal Society University Research Fellowship; NASA and UMBC/GEST for a Goddard Distinguished Visiting Fellowship in the Earth Sciences; The NERC, ESA and the EU for research support; and anonymous reviewers for their constructive criticisms and comments.

References

Anderson, D., The troposphere-stratosphere radiation-field at twilight - A spherical model, *Planet. Space Sci.*, *31*, 1517–1523, 1983.

Atkinson, R., D. Baulch, R. Cox, R. Hampson, J. Kerr, M. Rossi, and J. Troe, Evaluated kinetic and photochemical data for atmospheric chemistry: Supplement VI., IUPAC subcommittee on gas kinetic data evaluation for atmospheric chemistry, *J. Phys. Chem. Ref. Data*, *26*, 1329–1499, 1997.

Daly, R., *Atmospheric Data Analysis*, Cambridge Atmospheric and Space Science Series, Cambridge University Press, Cambridge, England, 1991.

DeMore, W. B., C. J. Howard, S. P. Sander, A. R. Ravishankara, D. M. Golden, C. E. Kolb, M. J. Hampson, R. F. Molina, and M. J. Kurylo, Chemical kinetics and photochemical data for use in stratospheric modeling, *JPL Publ. 97-4 12*, 1997.

Fisher, M., and D. Lary, Lagrangian 4-dimensional variational data assimilation of chemical-species, *Q. J. R. Meteorol. Soc.*, *121*, 1681-1704, 1995.

Khattatov, B. V., J. C. Gille, L. Lyjak, G. P. Brasseur, V. L. Dvortsov, A. E. Roche, and J. W. Waters, Assimilation of photochemically active species and a case analysis of UARS data, *J. Geophys. Res.*, *104*, 18,715–18,737, 1999.

Lary, D., and J. Pyle, Diffuse-radiation, twilight, and photochemistry: 1., *J. Atmos. Chem.*, *13*, 373–392, 1991a.

Lary, D., and J. Pyle, Diffuse-radiation, twilight, and photochemistry: 2., *J. Atmos. Chem.*, *13*, 393–406, 1991b.

Lary, D., M. Chipperfield, J. Pyle, W. Norton, and L. Riishojgaard, 3-dimensional tracer initialization and general diagnostics using equivalent pv latitude-potential-temperature coordinates, *Q. J. R. Meteorol. Soc.*, *121*, 187-210, 1995a.

Lary, D., M. Chipperfield, and R. Toumi, The potential impact of the reaction OH+ClO$\longrightarrow$HCl+ O_2 on polar ozone photochemistry, *J. Atmos. Chem.*, *21*, 61–79, 1995b.

Lary, D., Gas phase atmospheric bromine photochemistry, *J. Geophys. Res.*, *101*, 1505–1516, 1996.

Meier, R., D. Anderson, and M. Nicolet, Radiation-field in the troposphere and stratosphere from 240-1000 nm. 1., General-analysis, *Planet. Space Sci.*, *30*, 923–933, 1982.

Menard, R., and L. Chang, Assimilation of stratospheric chemical tracer observations using a kalman filter. part ii: chi(2)-validated results and analysis of variance and correlation dynamics, *Monthly Weather Review*, *128*, 2672–2686, 2000.

Menard, R., S. Cohn, and L. Chang, Assimilation of stratospheric chemical tracer observations using a kalman filter. part i: Formulation, *Monthly Weather Review*, *128*, 2654-2671, 2000.

Nicolet, M., R. Meier, and D. Anderson, Radiation-field in the troposphere and stratosphere .2. Numerical-analysis, *Planet. Space Sci.*, *30*, 935–983, 1982.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in Fortran - The Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, New York, 1992.

Stoer, J., and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

D. J. Lary, Data Assimilation Office, NASA Goddard Space Flight Centre (dlary@dao.gsfc.nasa.gov)

B. Khattatov, NCAR, Boulder, CO

H. Mussa, Department of Chemistry, University of Cambridge, England

(Received February 15, 2002; revised xx 2002;
accepted xx 2002.)

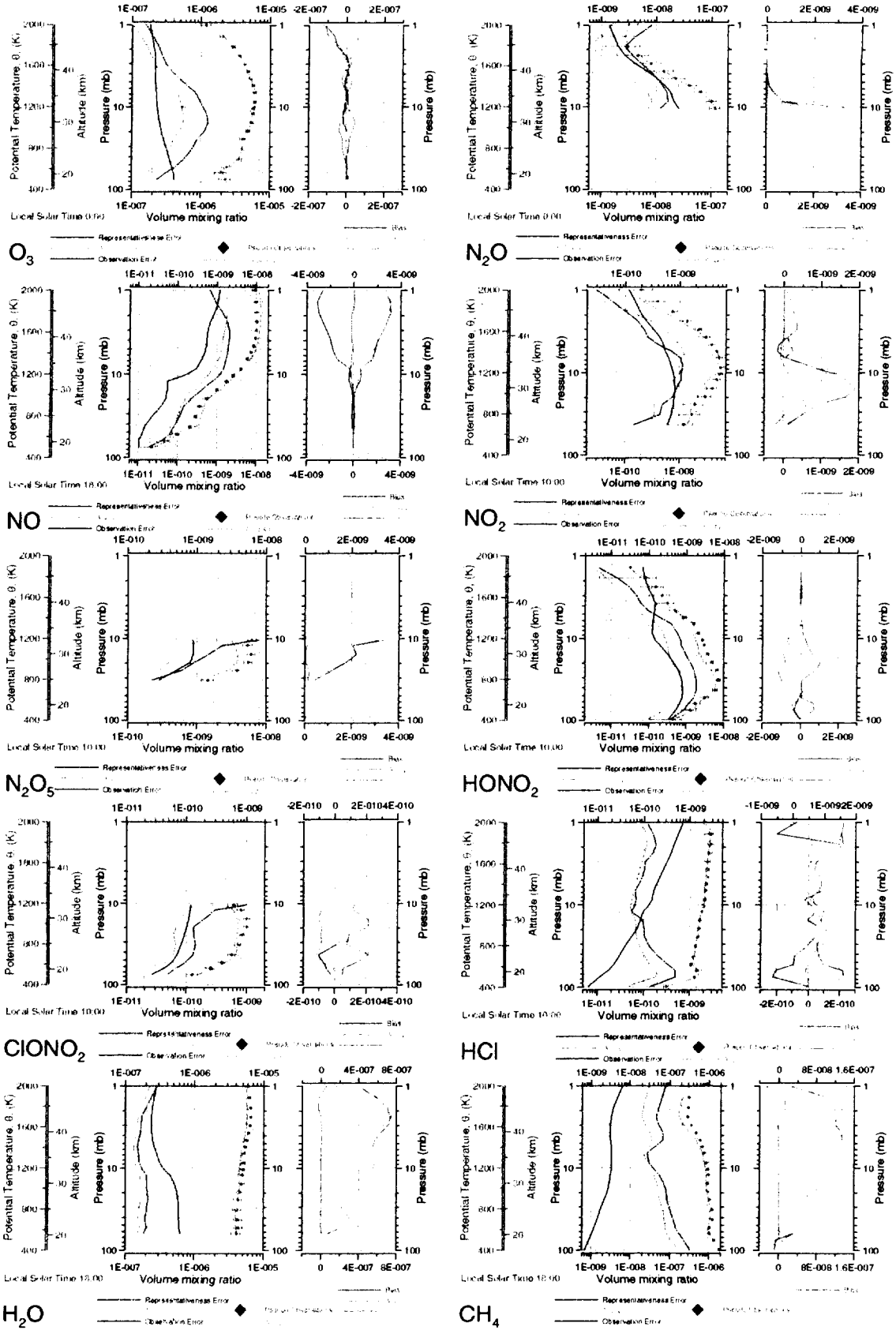

Figure 1. Vertical profiles of the chemical analyses produced by data assimilation for O_3, N_2O, NO, NO_2, N_2O_5, HNO_3, $ClONO_2$, HCl, H_2O and CH_4 overlaid with their pseudo observations. Shown on the same horizontal scale are vertical profiles of the representativeness error, the observation error and the analyses error. In a separate panel there are vertical profiles of the bias between the analyses and the pseudo observations, the analysis increment, (O-F), and (A-F).

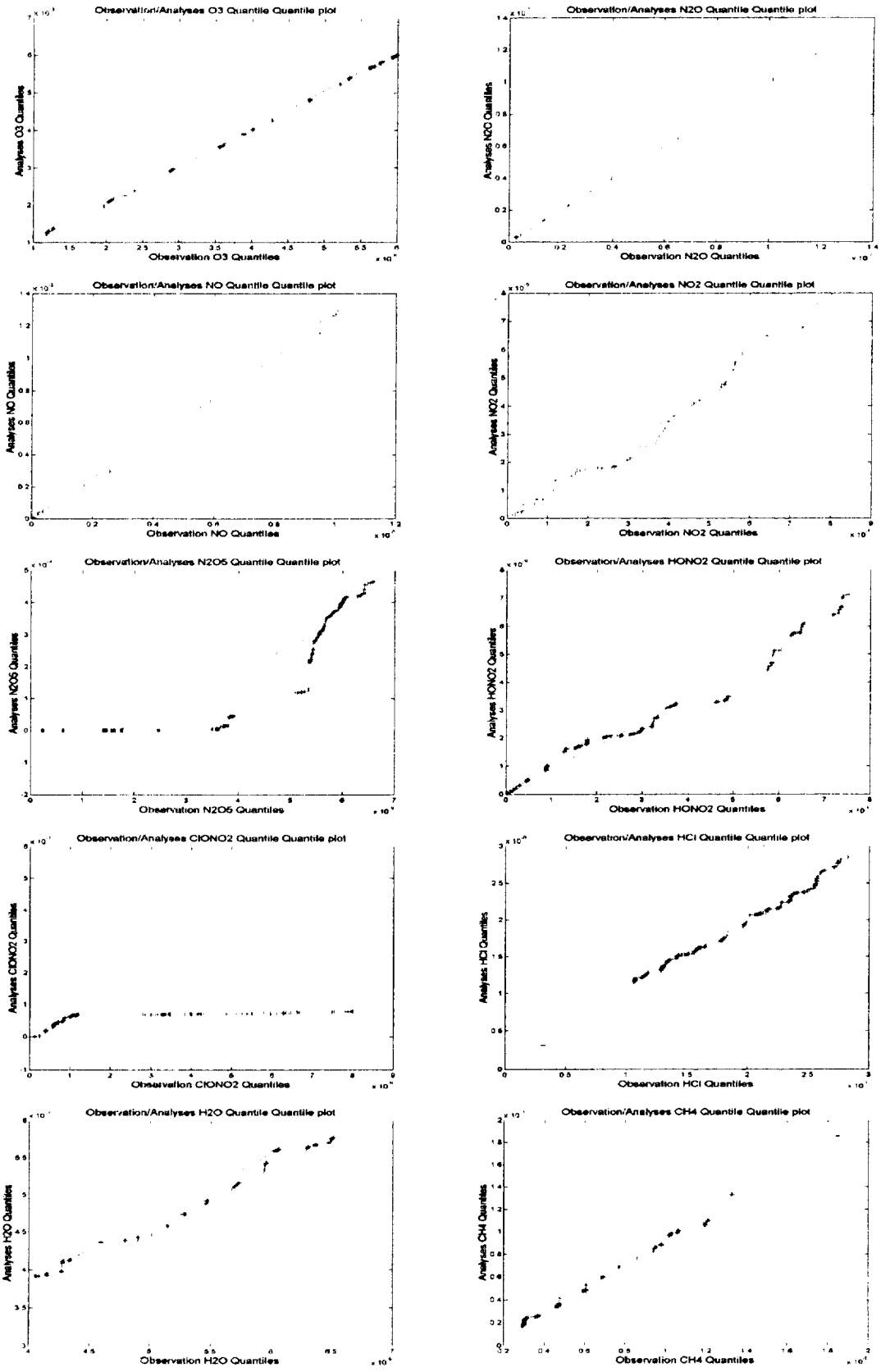

Figure 2. Quantile-Quantile Plots of observations against analyses for O_3, N_2O, NO, NO_2, N_2O_5, HNO_3, $ClONO_2$, HCl, H_2O and CH_4. A quantile-quantile plot is useful for determining whether two samples come from the same distribution (whether normally distributed or not). The quantile-quantile plot has three graphical elements. The pluses are the quantiles of each sample. The number of pluses is the number of data values in the smaller sample. The solid line joins the 25^{th} and 75^{th} percentiles of the

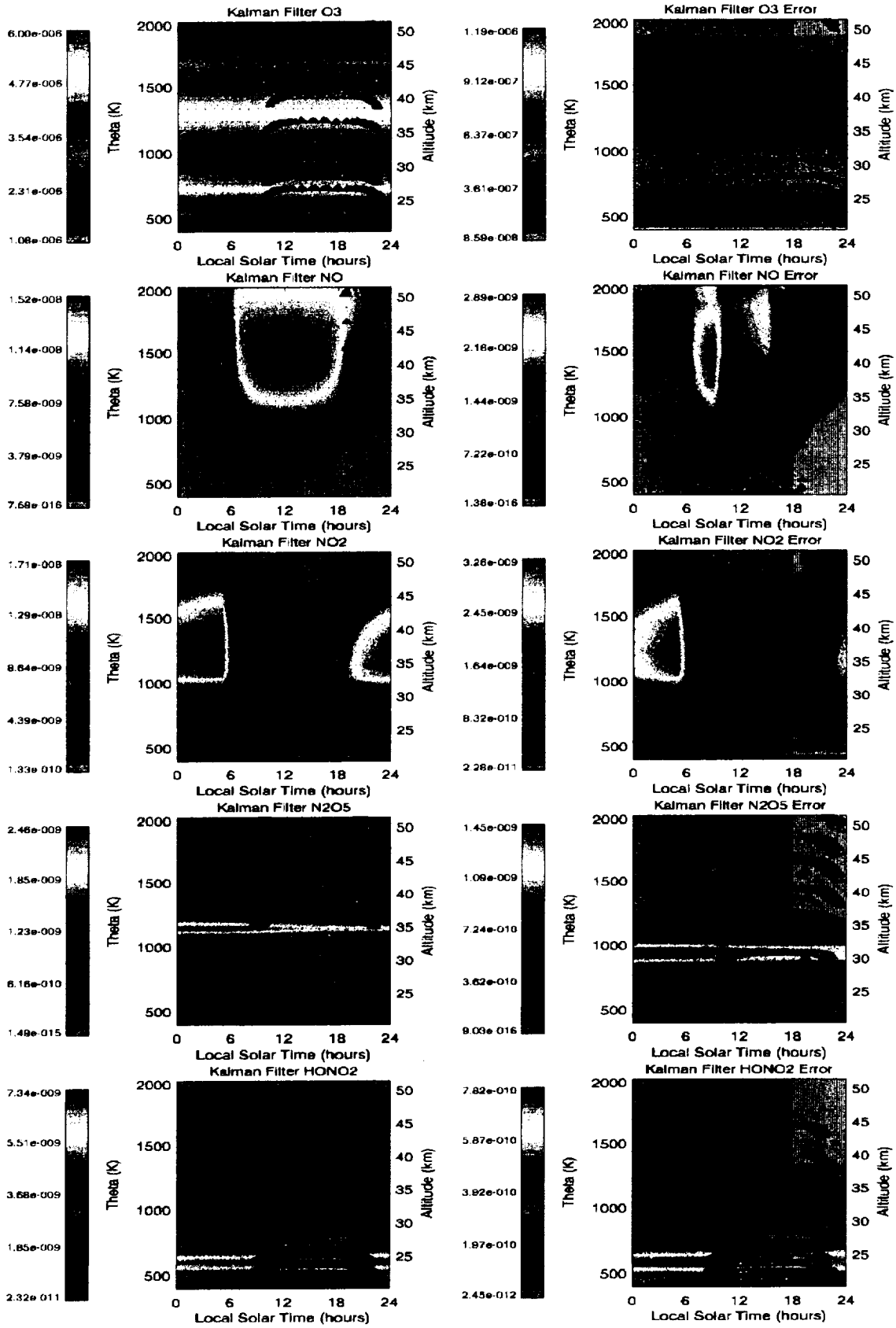

Figure 3. The left hand column shows one diurnal cycle of the chemical analyses for NO, NO_2, N_2O_5, and HNO_3 produced by data assimilation for a vertical profile at an equivalent PV latitude (ϕ_e) of 55°S consisting of 21 potential temperature (θ) levels spaced equally in $\log(\theta)$ between 400 K and 2000 K overlaid with the raw observations. The right hand column shows the corresponding analyses uncertainty overlaid with the observational uncertainty.

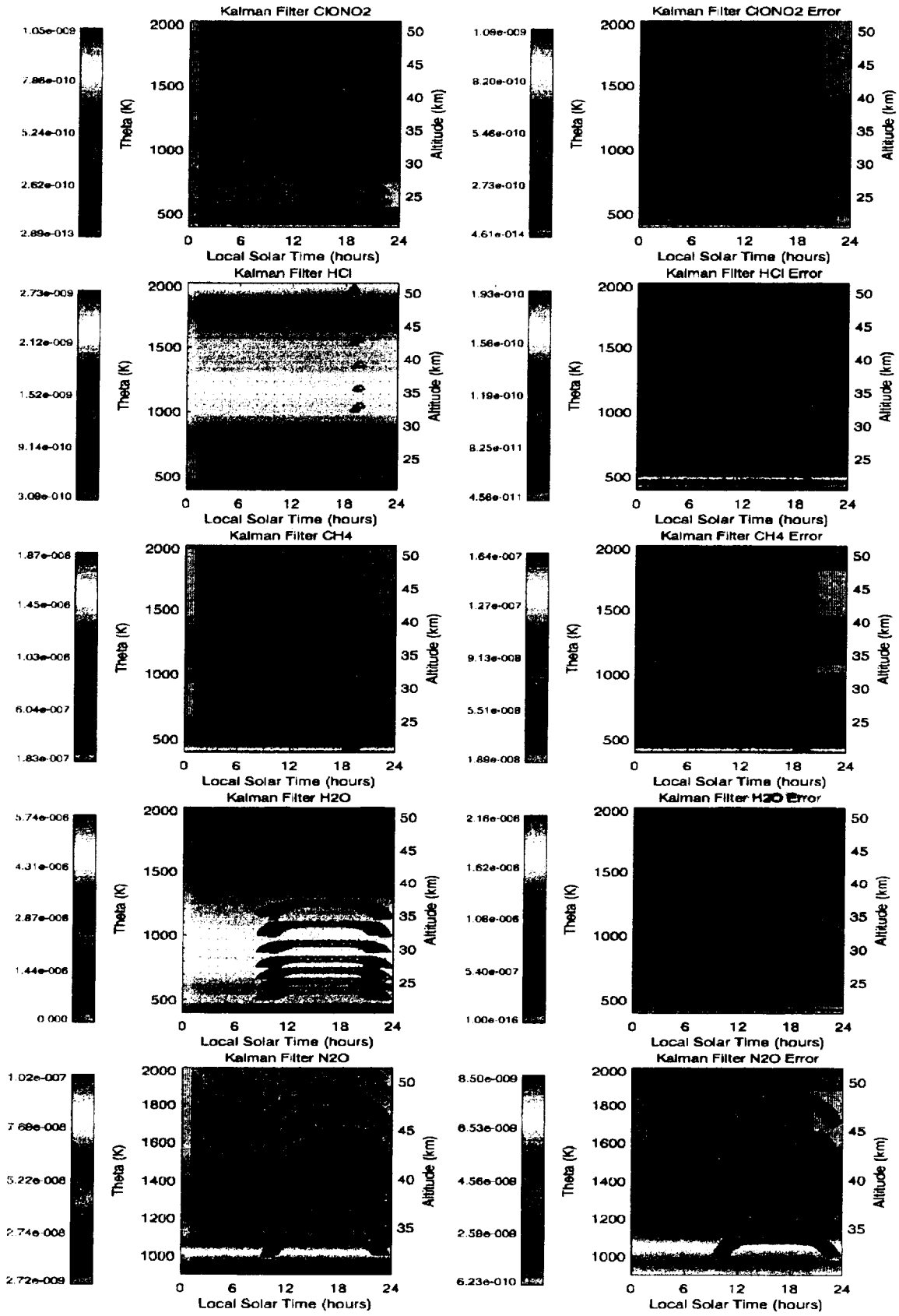

Figure 4. The left hand column shows one diurnal cycle of the chemical analyses for $ClONO_2$ HCl, CH_4, H_2O and N_2O produced by data assimilation for a vertical profile at an equivalent PV latitude (ϕ_e) of 55°S consisting of 21 potential temperature (θ) levels spaced equally in $\log(\theta)$ between 400 K and 2000 K overlaid with the raw observations. The right hand column shows the corresponding analyses uncertainty overlaid with the observational uncertainty.